УДК 004.912

О.В. Лозинська, М.В. Давидов, В.В. Пасічник
Національний університет «Львівська політехніка», Україна
вул. Ст. Бандери, 12, м. Львів, 79013

ВИКОРИСТАННЯ ЗВАЖЕНИХ АФФІКСНИХ КОНТЕКСТНО-ВІЛЬНИХ ГРАМАТИК ДЛЯ ЗМІШАНОГО СИНТАКСИЧНО-СЕМАНТИЧНОГО РОЗБОРУ РЕЧЕНЬ

O.V. Lozynska, M.V. Davydov, V.V. Pasichnyk
Lviv Polytechnic National University, Ukraine
S. Bandery str., 12, Lviv, 79013

USING OF WEIGHTED AFFIX CONTEXT–FREE GRAMMARS FOR MIXED SYNTACTIC-SEMANTIC PARSING SENTENCES

У статті розглянуто використання зважених афіксних контекстно-вільних граматик для змішаного синтаксично-семаничного розбору речень української мови. Представлена модифікація афіксної граматики над скінченною браткою (Affix grammar over a finite lattice, AGFL), яка додає семантичний атрибут і нову форму продукцій, яку названо «шаблонна продукція». Ця нова форма дає змогу створювати лаконічні і ефективні, з точки зору обчислень, продукції на основі онтологій. Вивчено нормальну форму шаблонних продукцій і запропоновано ефективний алгоритм для синтаксично-семантичного аналізу речень на їх основі. Проведено експерименти із використанням зважених афіксних контекстно-вільних граматик для синтаксично-семаничного розбору речень художнього тексту, які показали, що середній час розбору речень виявився практично лінійною функцією від кількості слів у них.

Ключові слова: синтаксично-семаничний розбір, зважена афіксна контекстно-вільна граматика, шаблонна продукція, онтологія.

The using of weighted affix context-free grammar for mixed syntactic–semantic parsing Ukrainian sentences is considered in the article. The modification of affix grammar over a finite lattice that adds semantical attribute and a new form of production called the “template production” is introduced. This new form helps to represent ontology-based productions in a short and computationally inexpensive way. The normal form of template production is studied, and effective algorithm for syntactic–semantic parsing sentences is proposed. The experiments with using of weighted affix context-free grammar for syntactic–semantic parsing of sentences from the test database of Ukrainian fiction literature are conducted. The growth of parsing time turned out to be almost linear function of the number of words in a sentence.

Keywords: syntactic–semantic parsing, weighted affix context-free grammar, template productions, ontology.

Вступ

Проблема автоматичного розбору текстів не є новою і все частіше виникає при створенні комп'ютерних додатків, які вирішують задачі машинного перекладу, пошуку інформації, класифікації документів, взаємодії між людьми та комп'ютером, моніторингу соціальних мереж тощо.

Задача синтаксично-семантичного розбору є складною задачею штучного інтелекту, оскільки її комплексне рішення вимагає побудови повної моделі людського знання. Хоча такі моделі в даний час розробляються [1], досі немає повноцінного рішення.

Для синтаксично-семантичного аналізу запропоновано підхід з використанням зваженої афіксної контекстно-вільної граматики (weighted affix context-free grammar, WACFG), яка є модифікацією відомої афіксної граматики над скінченною браткою (AGFL). WACFG використовує переваги ймовірнісної контекстно-вільної граматики (PCFG) [2] та афіксної граматики над скінченною браткою, розробленою К. Костером [3]. Відомі зважені та стохастичні граматики також застосовуються [4], але підхід, що базується на вагах, є більш гнучким.

Постановка проблеми

Основною метою цієї статті є розроблення підходу для ефективного подання зваженої афіксної контекстно-вільної граматики за допомогою спеціальної форми «шаблонна продукція».

Ця стаття описує метод, у якому семантичний аналіз інтегрований в алгоритм синтаксичного аналізу. Цей підхід допомагає зменшити кількість проміжних конструкцій, які необхідно розглянути. Це особливо важливо для флективних мов, таких як українська та інші слов'янські мови.

Проблема полягає в розробленні ефективних методів інтеграції семантичних атрибутів у продукції зваженої афіксної контекстно-вільної граматики та розробленні ефективного алгоритму розбору речень.

Аналіз останніх досліджень та публікацій

Проблема синтаксичного розбору речень вивчається протягом тривалого часу. Серед багатьох методів розбору речень підхід на основі породжувальних граматик, запропонований Н. Хомским [5], є одним з найбільш вивчених. Розширені афіксні граматики (EAG) [6] та ймовірнісні контекстно-вільні граматики [2] є породжувальними розширеними фундаментальними граматиками, що широко використовуються в лінгвістичних програмах у даний час.

Афіксні граматики, які належать до сімейства дворівневих граматик, є підмножиною розширених граматик. Продукції афіксної граматики є продукціями, які розширені атрибутами. Домен атрибутів визначається метаграматикою.

Ефективність афіксних граматик над скінченною граткою та їхнє застосування у алгоритмі розбору речень були доведені К. Костером [3]. Розширення AGFL, які базуються на ймовірностях, також вивчали Т. Сміт та Дж. Клірі [7].

Запропонований авторами підхід на основі зваженої афіксної граматики над скінченною граткою є близьким до методу, введеного К. Костером. Однак ми формулюємо цю граматику та продукції по-іншому, що дає змогу використовувати коротку форму продукцій та компактний алгоритм розбору речень.

Метод змішаного синтаксично-семантичного розбору речень з використанням WACFG

Зважена афіксна граматика над скінченною граткою. Для змішаного синтаксично-семантичного розбору речень розроблено новий парсер, який використовує зважену афіксну граматику над скінченною граткою. Ця граматика розширює символи породжувальної граматики афіксами, які можуть бути використані для зменшення кількості продукцій, необхідних для опису мови. Подане в статті означення афіксної граматики над скінченною граткою дещо відрізняється від запропонованого К. Костером, але воно має таку саму ідею. Це нове означення було використано для доведення того, що деякі правила перетворення можуть бути застосовані до граматики для прискорення процесу синтаксичного аналізу.

Зважена афіксна граматика над скінченною граткою G визначається як кортеж (T,V,S,D,P), де T являє собою множину всіх термінальних символів, V являє собою множину всіх символів, $S \in V \setminus T$ – це початковий символ, який являє собою множину афіксних доменів, що не перетинаються; кожен домен D_i являє собою множину афіксів $A(D_i)$; P – це множина шаблонних і регулярних продукцій.

Регулярні продукції мають форму $\left\langle \left(V \times 2^A\right)^* \overset{w}{\to} \left(V \times 2^A\right)^* \right\rangle$, де A – це множина всіх

афіксів з $A = A(D) = \bigcup_{D_j \in D} A(D_j)$, 2^A позначає потужність множини A, і $(V \times 2^A)^*$ позначає всі непусті рядки атрибутивних символів $s_1 s_2 \ldots s_k$, з $k > 0$, $s_j = (v_j, A_j)$, $(v_j, A_j) \in V \times 2^A$; $w \in R^+$ є мультиплікативна вага продукції. Символ ваги можна пропустити, коли він дорівнює 1.

Термінальні символи $t_i \in T$ не мають атрибутів. Вони зазвичай являють собою слова речень після синтаксичного аналізу. Наприклад, слово «лікар» може бути чоловічим чи жіночим іменником однини, перш ніж він буде відомим з контексту. Згідно з термінами породжувальної граматики, це може бути написано таким чином:

$$(noun,\{a_{FEMALE}, a_{SINGULAR}, a_{DOCTOR}\}) \to (\text{лікар}, \varnothing),$$

$$(noun,\{a_{MALE}, a_{SINGULAR}, a_{DOCTOR}\}) \to (\text{лікар}, \varnothing).$$

Альтернативна форма може бути записана таким чином: $(noun,\{a_{FEMALE}, a_{MALE}, a_{SINGULAR}, a_{DOCTOR}\}) \to (\text{лікар}, \varnothing)$. Вона являє собою обидва випадки, наведені вище. Продукції, які генерують термінальні символи, додаються морфологічним аналізатором. Якщо якесь слово є омографом, морфологічний аналізатор генерує одну продукцію для кожного значення слова. Вага кожної продукції відображає допустимість цього значення в аналізованому контексті.

У наведеному вище прикладі $a_{FEMALE}, a_{MALE}, a_{SINGULAR}$ – це граматичні атрибути, $a_{STUDENT}$ є семантичним атрибутом. Семантичні атрибути – це елементи домену D_{SEM}.

Забезпечення регулярних продукцій для всіх можливих комбінацій афіксів може бути неефективним. Для цього вводиться шаблонна форма продукцій. Ця форма розроблена для підвищення обчислювальної ефективності опрацювання мови.

Шаблона продукція має форму $(v_1, D_{inh\,1}, A_{set\,1}) \ldots (v_k, D_{inh\,k}, A_{set\,k}) \overset{w}{\to} (v'_1, D_{uni\,1}, A_{req\,1}), \ldots (v'_m, D_{uni\,m}, A_{req\,m})$, де $D_{inh\,1}, D_{inh\,2}, \ldots, D_{inh\,k} \subset D$ є доменами, афікси яких успадковуються із символів $v_1, v_2, \ldots, v_k$; $D_{uni\,1}, D_{uni\,2}, \ldots, D_{uni\,m} \subset D$ є доменами, афікси яких повинні бути загальними для символів v'_1, v'_2, ..., v'_m; $A_{set\,1}$, $A_{set\,2}$, …, $A_{set\,k} \subset A$ є додатковими афіксами для символів у лівій частині продукції, та $A_{req\,1}$, $A_{req\,2}$, …., $A_{req\,m} \subset A$ – це необхідні афікси для символів у правій частині продукції; w – це мультиплікативна вага продукції.

Форма шаблона за означенням еквівалентна множині регулярних продукцій. Розглянемо наступний шаблон і регулярні продукції (1) і (2):

$$q = \left\langle (v_1, D_{inh\,1}, A_{set\,1}) \ldots (v_k, D_{inh\,k}, A_{set\,k}) \overset{w}{\to} (v'_1, D_{uni\,1}, A_{req\,1}) \ldots (v'_m, D_{uni\,m}, A_{req\,m}) \right\rangle, \qquad (1)$$

$$p = \left\langle (v_1, A_1) \ldots (v_k, A_k) \overset{w}{\to} (v'_1, A'_1) \ldots (v'_m, A'_m) \right\rangle. \qquad (2)$$

Нехай, як $A_{uni}(p,q)$ позначимо перетин всіх атрибутів, які повинні бути однаковими у правій частині регулярної продукції p, щоб відповідати шаблонній продукції q:

$$A_{uni}(p,q) = \bigcap_{i=1}^{m} \left(A'_i \cup \overline{A}(D_{uni\,i}) \right) \cap A(D_{uni\,1..m}),$$

$$\overline{A}(D_{uni\,i}) = A \;\; A(D_{uni\,i}),\; D_{uni\,1..m} = D_1 \cup D_2 \cup \ldots \cup D_m .$$

Говорять, що регулярна продукція p відповідає шаблонній продукції q, якщо виконуються вимоги R1-R3:

R1. $(\forall j \in 1\ldots n) D_j \in D_{uni\,1..m} \Rightarrow A_{uni}(p,q) \cap A(D_j) \neq \varnothing$;

R2. $(\forall i \in 1\ldots m) A_{req\,i} \subset A'_i$;

R3. $(\forall i \in 1\ldots k) A_i = A_{set\,i} \cup (A_{uni}(p,q) \cap A(D_{inh\,i}))$.

Вимога R1 гарантує, що для кожного уніфікованого домену є, принаймні один, загальний афікс. Вимога R2 описує, як обробляються обов'язкові атрибути, а вимога R3 вказує на те, як отримуються атрибути символів у лівій частині продукції.

Наприклад, український еквівалент англійської назви фрази «GREEN LEAVES OF THE TREE» є «ЗЕЛЕНЕ ЛИСТЯ ДЕРЕВА». У цьому іменниковому словосполученні відмінок, стать і число прикметника (ЗЕЛЕНЕ) координуються за відмінком, статтю та числом першого іменника (ЛИСТЯ), а відмінок другого іменника (ДЕРЕВА) є ЗНАХІДНИМ. Семантичний атрибут для усього словосполучення взятий зі слова «ДЕРЕВА». Шаблонна продукція для цього словосполучення українською мовою:

$$(NP,\{D_{GENDER},D_{NUMBER},D_{CASE},D_{SEM}\},\varnothing) \to (ADJ,\{D_{GENDER},D_{NUMBER},D_{CASE}\},\varnothing)$$
$$(NP,\{D_{GENDER},D_{NUMBER},D_{CASE},D_{SEM}\},\varnothing)(NP,\varnothing,\{a_{GENITIVE}\}),$$

і англійський еквівалент:

$$(NP,\{D_{NUMBER},D_{SEM}\},\varnothing) \to (ADJ,\varnothing,\varnothing)(NP,\{D_{NUMBER},D_{SEM}\},\varnothing)(prep,\varnothing,\{a_{OF}\})(NP,\varnothing,\varnothing),$$

де NP стоїть біля іменникового словосполучення, ADJ стоїть біля прикметника, D_{GENDER}, D_{NUMBER}, D_{CASE}, D_{SEM} – це домени для позначення статі, числа, відмінка та семантичних афіксів, відповідно.

Нормальна форма шаблонних продукцій. Довжина правої частини продукції називається її рангом. Ефективний синтаксичний аналіз речень за допомогою породжувальних граматик може бути досягнутий, коли граматика знаходиться в нормальній формі Хомського (CNF) – формі, яка гарантує, що всі продукції граматики мають ранг не більше 2. Шаблонні продукції також мають бути перетворені у форму, яка має не більше двох символів у правій частині. Це перетворення виконується шляхом застосування кроків спрощення для всіх продукцій, які мають ранг більше 2. На кожному кроці береться одна шаблонна продукція з рангом $m > 2$ і утворюються дві шаблонні продукції – одна з рангом 2 і одна з рангом $m-1$. Процес зупиняється, коли немає більше продукцій із рангом 3 та вище.

На кроці спрощення береться одна шаблонна продукція q форми (1) і утворюються 2 шаблонні продукції:

$$q_1 = \left\langle (v_1, D_{inh\,1}, A_{set\,1})\ldots(v_k, D_{inh\,k}, A_{set\,k}) \overset{w}{\to} (v'_1, D_{uni\,1}, A_{req\,1})(v'_{2..m}, D_{uni\,2..m}, \varnothing) \right\rangle,$$

$$q_2 = \left\langle (v'_{2..m}, D_{uni\,2..m}, \varnothing) \overset{1.0}{\to} (v'_2, D_{uni\,2}, A_{req\,2})\ldots(v'_m, D_{uni\,m}, A_{req\,m}) \right\rangle,$$

де $D_{uni\,2..m} = D_{uni\,2} \cup D_{uni\,3} \cup \ldots \cup D_{uni\,m}$ і $v'_{2..m}$ – це новий нетермінальний символ.

Алгоритм синтаксичного розбору речень

Проблема синтаксичного розбору речень сформульована як проблема пошуку послідовності продукцій, що мають максимальну вагу, і можуть бути застосовані послідовно до деякого початкового атрибутивного символу (S, A_s) для створення

заданої послідовності терміналів $t_1 t_2 \ldots t_n$. Вага послідовності розраховується як добуток ваг усіх збережених продукцій.

Блок-схема алгоритму розбору показана на рис. 1.

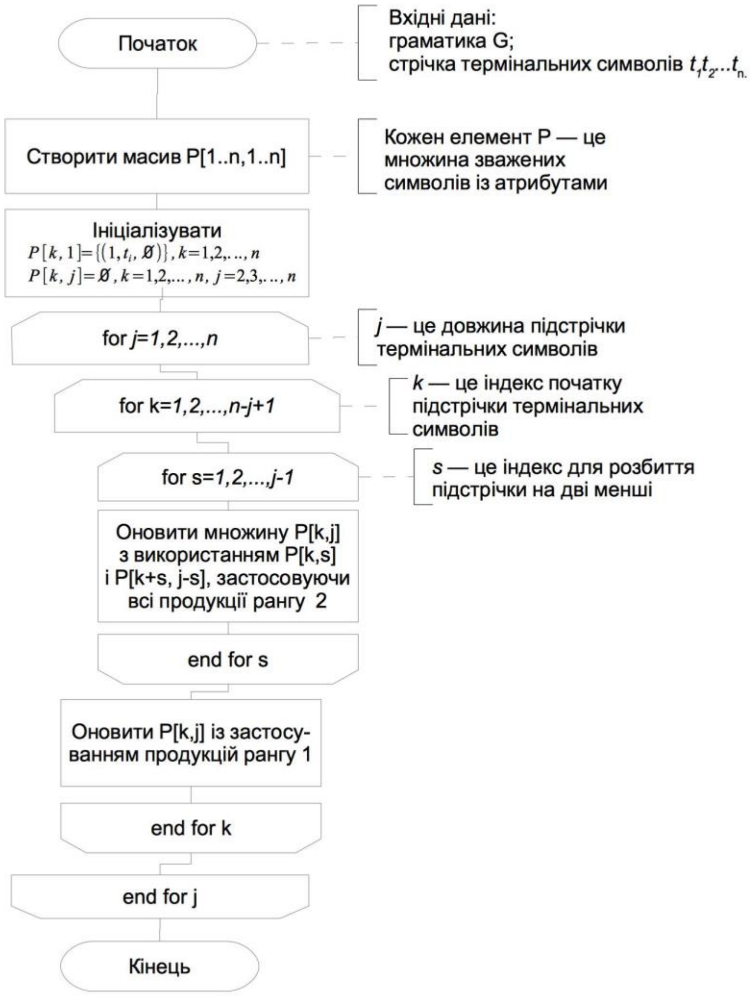


Рис. 1. Блок-схема алгоритму синтаксично-семантичного розбору речень

Наведений вище алгоритм використовує внутрішню процедуру для оновлення множини зважених атрибутивних символів Q з множини можливих лівих символів L та множини можливих правих символів R із застосуванням рангу 2. Блок-схема цієї процедури зображена на Рис. 2.

Якщо в правій частині продукції міститься лише один символ, вага продукції не повинна перевищувати 1, щоб уникнути циклічних продукцій, що збільшують вагу нетермінальних символів під час процедури розбору знизу-вверх.

Розроблений алгоритм розбору речень побудований в основному на ймовірнісному CYK-алгоритмі. Головна відмінність полягає в тому, що символи порівнюються не тільки за вагою, а й із множиною афіксів. Алгоритм використовує поняття зваженого атрибутивного символу – це кортеж (w, v, A_v), що містить вагу w, символ v та множину афіксів $A_v \subset A(D)$. Зважений атрибутивний символ (w_1, v_1, A_1) домінує над зваженим атрибутивним символом, якщо $w_1 \geq w_2$, $v_1 = v_2$, і $A_2 \subset A_1$.

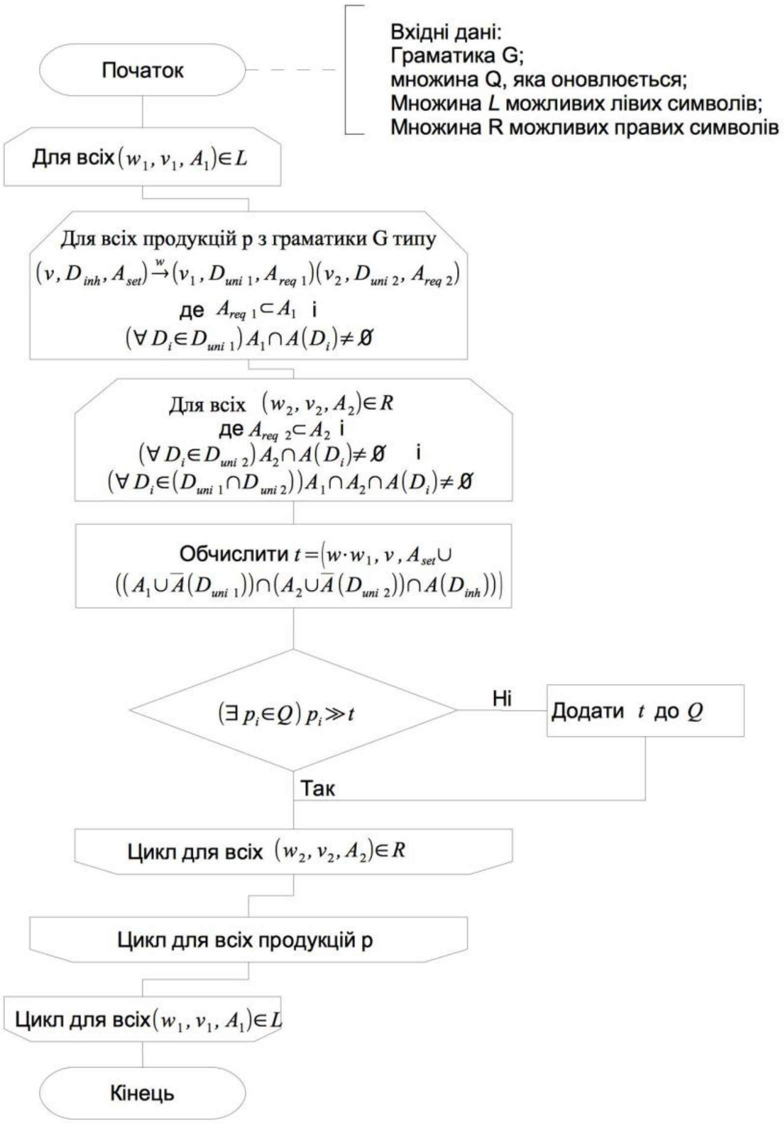


Рис. 2. Блок-схема процедури оновлення множини зважених атрибутивних символів Q

У найгіршому випадку, обчислювальна складність запропонованого алгоритму складає $O(n^3 \cdot m_p^3 \cdot m_r)$, де n – довжина вхідного рядка терміналів, m_p – це максимальна кількість комбінацій символів граматики та атрибутів, з яких можна одержати той самий рядок терміналів (це значення можна розглядати як неоднозначність мови, що аналізується), і m_r – це максимальна кількість продукцій, які мають однаковий початковий нетермінальний символ у правій частині.

Аналіз результатів

Алгоритм розбору речень реалізований у проекті відкритого програмного забезпечення UkrParser [8]. Цей проект містить класи для морфологічного та синтаксичного аналізу речень. Обчислювальна ефективність розробленого алгоритму перевірена на базі даних 500 речень з оповідань Михайла Коцюбинського. Середній час розбору речень залежно від довжини речення зображений на Рисунку

3. Ці результати були отримані на комп'ютері з процесором 2,4 ГГц Intel Core i5. Зростання часу синтаксичного аналізу виявилося практично лінійним.

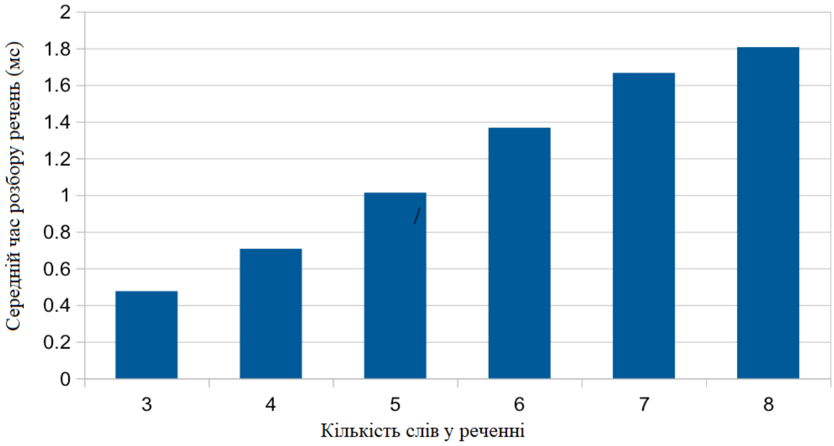


Рис. 3. Середній час розбору речення в мілісекундах залежно від довжини речення

Розроблений підхід для змішаного синтаксично-семантичного аналізу речень був використаний для аналізу та перекладу анотованої української жестової мови та української словесної мови [9], де переклад на основі синтаксично-семантичного аналізатора, який використовує продукції, згенеровані з використанням онтології, показав кращі результати ніж аналізатор, який використовує тільки синтаксичні продукції на 25% (90% правильних перекладів у порівнянні з 65% правильних перекладів, отриманих при використанні лише синтаксичних продукцій).

Висновки

У статті наведено ефективний алгоритм розбору речень за допомогою зваженої афіксної контекстно-вільної граматики з семантичними атрибутами. Розроблений алгоритм використовує нормальну форму «шаблонних продукцій». Алгоритм має кубічну складність, але на практиці зростання часу обчислення виявилося майже лінійним відносно кількості слів у реченні. Отримані дерева синтаксично-семантичного розбору речення, мають більше семантичних атрибутів, ніж дерева синтаксичного розбору, отримані за допомогою синтаксичного аналізатора. Додаткові обчислювальні затрати для цього невеликі, оскільки в граматику включені лише гіперніми слів, що містяться в реченні та відповідних словосполученнях.

Подальші дослідження будуть зосереджені на оптимальному розподілі ваги та автоматичному утворенню продукцій, специфічних для конкретної предметної області.

Література

1. Najmi E. ConceptOnto: An upper ontology based on ConceptNet / E. Najmi, K. Hashmi, Z. Malik, A. Rezgui, H.U. Khanz // Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), November 10-13, Doha, Qatar. – 2014. – P. 366-372. DOI: 10.1109/AICCSA.2014.7073222.
2. Eddy S.R. RNA sequence analysis using covariance models / S.R. Eddy, R Durbin // Nucleic Acids Research. – 1994. – Vol. 22. – № 11. – P. 2079–2088.
3. Koster C.H.A. Affix Grammars for natural languages / C.H.A. Koster // In: Attribute Grammars, Applications and Systems, International Summer School SAGA, Lecture Notes in Computer Science, Prague, Czechoslovakia. – 1991. – Vol. 545. – P. 469-484.
4. Smith N.A. Weighted and Probabilistic Context-Free Grammars Are Equally Expressive / N.A. Smith, M. Johnson // Computational Linguistics. – 2007. – Vol. 33. – № 4. – P. 477-491. DOI:10.1162/coli.2007.
5. Chomsky N. Three models for the description of language / N. Chomsky // IRE Transactions on Information Theory 2 (3). – 1956. – P. 113–124. DOI:10.1109/TIT.1956.1056813.

6. Oostdijk N. An Extended Affix Grammar for the English Noun Phrase / N. Oostdijk // In: Jan Aarts and Wim Meijs (eds), Corpus Linguistics. Recent Developments in the Use of Computer Corpora in English Language Research, Amsterdam: Rodopi. – 1984.
7. Smith T.C. Probabilistic Unification Grammars / T.C. Smith, J.G. Cleary // In Australasian Natural Language Processing Summer Workshop. – 1997. – P. 25-32.
8. Проект «UkrParser» [Електр. ресурс]. – Режим доступу: https://github.com/mdavydov/UkrParser.
9. Davydov M. Spoken and sign language processing using grammatically augmented ontology / M. Davydov, O. Lozynska // Applied Computer Science. – Poland, 2015. – Vol. 11. – No 2. – P. 29–42.

Literatura

1. Najmi E. ConceptOnto: An upper ontology based on ConceptNet / E. Najmi, K. Hashmi, Z. Malik, A. Rezgui, H.U. Khanz // Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), November 10-13, Doha, Qatar. – 2014. – P. 366-372. DOI: 10.1109/AICCSA.2014.7073222.
2. Eddy S.R. RNA sequence analysis using covariance models / S.R. Eddy, R Durbin // Nucleic Acids Research. – 1994. – Vol. 22. – № 11. – P. 2079–2088.
3. Koster C.H.A. Affix Grammars for natural languages / C.H.A. Koster // In: Attribute Grammars, Applications and Systems, International Summer School SAGA, Lecture Notes in Computer Science, Prague, Czechoslovakia. – 1991. – Vol. 545. – P. 469-484.
4. Smith N.A. Weighted and Probabilistic Context-Free Grammars Are Equally Expressive / N.A. Smith, M. Johnson // Computational Linguistics. – 2007. – Vol. 33. – № 4. – P. 477-491. DOI:10.1162/coli.2007.
5. Chomsky N. Three models for the description of language / N. Chomsky // IRE Transactions on Information Theory 2 (3). – 1956. – P. 113–124. DOI:10.1109/TIT.1956.1056813.
6. Oostdijk N. An Extended Affix Grammar for the English Noun Phrase / N. Oostdijk // In: Jan Aarts and Wim Meijs (eds), Corpus Linguistics. Recent Developments in the Use of Computer Corpora in English Language Research, Amsterdam: Rodopi. – 1984.
7. Smith T.C. Probabilistic Unification Grammars / T.C. Smith, J.G. Cleary // In Australasian Natural Language Processing Summer Workshop. – 1997. – P. 25-32.
8. Proekt «UkrParser» [Elektr. resurs]. – Rezhym dostupu: https://github.com/mdavydov/UkrParser.
9. Davydov M. Spoken and sign language processing using grammatically augmented ontology / M. Davydov, O. Lozynska // Applied Computer Science. – Poland, 2015. – Vol. 11. – No 2. – P. 29–42.

RESUME

O.V. Lozynska, M.V. Davydov, V.V. Pasichnyk

Using of weighted affix context-free grammars for mixed syntactic–semantic parsing sentences

The problem of automatic text parsing is not new and more and more arises when creating the computer applications that solve machine translation tasks, information search, document classification, interaction between people and computer, social network monitoring, etc.

The task of increasing efficiency of affix grammars over a finite lattice is considered. For this purpose, the modification of affix grammar over a finite lattice that adds semantical attribute and a new form of production called the “template production” is implemented. This new form helps to represent ontology-based productions in a short and computationally inexpensive way. The normal form of template production is studied, and effective algorithm for syntactic–semantic parsing sentences is proposed. The using of weighted affix context-free grammar for mixed syntactic–semantic parsing Ukrainian sentences is introduced.

A new algorithm for mixed syntactic–semantic parsing sentence and a new procedure for updating set of weighted attributes symbols Q are developed. The experiments with using of weighted affix context-free grammar for syntactic–semantic parsing of sentences from the test database of Ukrainian fiction literature are conducted. The growth of parsing time turned out to be almost linear function of the number of words in a sentence.

Надійшла до редакції 03.11.2017

